AMENDMENTS TO THE CLAIMS:

1. (Previously presented) A method of improving at least one of speed and efficiency when executing a linear algebra subroutine on a computer having a memory hierarchical structure including at least one cache, said method comprising:

determining, based on sizes, for a level 3 matrix multiplication processing, which matrix will have data for a submatrix block residing in a lower level cache of said computer and which two matrices will have data for submatrix blocks residing in at least one higher level cache or a memory; and

streaming data from said selected two matrices, for executing said level 3 matrix multiplication processing, so that said submatrix block residing in said lower level cache remains resident in said lower level cache.

2. (Previously presented) The method of claim 1, wherein said lower level cache comprises an L1 cache and said higher level cache comprises an L2 cache.

3. (Previously presented) The method of claim 1, wherein said determining said matrix to be stored in said lower level cache comprises determining which of the three matrices has a smallest size.

4-5. (Canceled)

6. (Previously presented) The method of claim 2, wherein data for said second matrix and said third matrix streams into said L1 cache from said L2 cache such that said

data from said second matrix and said third matrix streams in a vector format into said L1 cache.

7. (Previously presented) The method of claim 1, wherein said linear algebra subroutine comprises a substitute of a subroutine from LAPACK (Linear Algebra PACKage).

8. (Previously presented) The method of claim 7, wherein said substitute subroutine comprises a BLAS (Basic Linear Algebra Subroutine) Level 3 routine or a BLAS Level 3 kernel routine.

9. (Previously presented) An apparatus, comprising:

a memory system to store matrix data for a level 3 matrix multiplication processing using data from a first matrix, a second matrix, and a third matrix, said memory system including at least one cache; and

a processor to perform said level 3 matrix multiplication processing, wherein data from one of said first matrix, said second matrix, and said third matrix is stored as a submatrix block resident in a lower level cache in a matrix format and data from a remaining two matrices is stored as submatrix blocks in said memory system at a level in said memory system higher than said lower level cache,

said processor preliminarily selecting, based on sizes, which matrix will have said submatrix block stored in said lower level cache and which said two matrices will have submatrix blocks stored in said higher level,

said data from said selected two matrices being streamed through said lower level cache into said processor, as required by said level 3 matrix multiplication processing, so that said submatrix block stored in said lower level cache remains resident in said lower level cache.

10. (Previously presented) The apparatus of claim 9, wherein said processor selects a smallest of said first, second, and third matrices to be said matrix to have data residing in said first level cache.

11. (Previously presented) The apparatus of claim 9, wherein said level 3 matrix multiplication comprises one or more subroutines substitute to ~~a~~ subroutines from LAPACK (Linear Algebra PACKage).

12. (Previously presented) The apparatus of claim 11, wherein said substitute subroutine comprises a BLAS (Basic Linear Algebra Subroutine) Level 3 routine or a BLAS Level 3 kernel routine.

13. (Canceled)

14. (Previously presented) A machine-readable storage medium tangibly embodying a program of machine-readable instructions executable by a digital processing apparatus to perform a method of improving at least one of a speed and an efficiency of executing a level 3 matrix multiplication processing on a computer having at least one lower level

cache and one or more higher level caches or other higher level memory devices, said method comprising:

selecting, based on sizes, which matrix will have submatrix block of data residing in said lower level cache and which two of three matrices will have submatrix blocks of data residing in at least one said higher level cache or memory; and

streaming data from said two selected matrices, for said level 3 matrix multiplication processing, so that said submatrix block residing in said lower level cache remains resident.

15. (Previously presented) The machine-readable storage medium of claim 14, wherein a smallest of said three matrices is selected as said matrix to have data to reside in said lower level cache.

16. (Previously presented) The machine-readable storage medium of claim 14, wherein said level 3 matrix multiplication processing comprises a substitute of a subroutine from LAPACK (Linear Algebra PACKage).

17. (Previously presented) The machine-readable storage medium of claim 16, wherein said substitute subroutine comprises a BLAS (Basic Linear Algebra Subroutine) Level 3 routine or a BLAS Level 3 kernel routine.

18. (Previously presented) The machine-readable storage medium of claim 14, wherein said lower level cache comprises an L1 cache and data for said second matrix and said third matrix streams from said higher level such that said data from one of said second matrix and said third matrix streams in a vector format through said L1 cache.

19. (Previously presented) A method of providing a service involving at least one of solving and applying a scientific/engineering problem, said method comprising at least one of:

using a linear algebra software package that performs a level 3 matrix multiplication processing, said software package comprising:

examining a size of each of three matrices involved in said level 3 matrix multiplication processing to select a smallest of said three matrices to have a block of submatrix data residing in an L1 cache and two remaining matrices to have data streamed from a higher level of memory; and

executing said level 3 matrix multiplication processing by said streaming of data from said two selected matrices, said streaming occurring so that said submatrix block residing in said L1 cache remains resident in said L1 cache;

providing a consultation for solving a scientific/engineering problem using said linear algebra software package;

transmitting a result of said linear algebra software package on at least one of a network, a signal-bearing medium containing machine-readable data representing said result, and a printed version representing said result; and

receiving a result of said linear algebra software package on at least one of a network, a signal-bearing medium containing machine-readable data representing said result, and a printed version representing said result.

20. (Canceled)

21. (Previously presented) The method of claim 1, said computer having M levels of caches and a main memory, said method further comprising:

selecting, from a plurality of six kernels, two kernels optimal to use for executing said level 3 matrix multiplication processing as data streams from different levels of said M levels of cache, such that said processor switches back and forth between said two selected kernels as steaming data traverses said different levels of cache.

22. (Previously presented) The machine-readable storage medium of claim 14, said memory system including M levels of caches and a main memory, wherein said processor initially selects, from a plurality of six kernels, two kernels optimal to use for executing said level 3 matrix multiplication processing as data streams from different levels of said M levels of cache, such that said processor switches back and forth between said two selected kernels as steaming data traverses said different levels of cache.

23. (Previously presented) The apparatus of claim 9, wherein said computer comprises M levels of caches and a main memory, said processor further preliminarily selecting, from a plurality of six kernels, two kernels optimal to use for executing said level 3 matrix multiplication processing as data streams from different levels of said M levels of cache, such that said processor switches back and forth between said two selected kernels as steaming data traverses said different levels of cache.